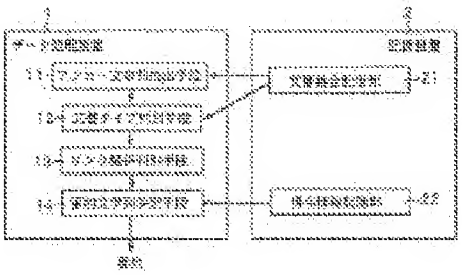
DOCUMENT SUMMARIZING SYSTEM AND DOCUMENT SUMMARIZING METHOD

Publication number: JP2002163276 (A)
Publication date: 2002-06-07
Inventor(s): AKAMINE SUSUMU; SUGIURA ATSUSHI +
Applicant(s): NEC CORP +
Classification:
- international: ***G06F12/00; G06F17/30;*** **G06F12/00; G06F17/30;** (IPC1-7): G06F12/00; G06F17/30
- European:
Application number: JP20000358808 20001127
Priority number(s): JP20000358808 20001127

Abstract of **JP 2002163276 (A)**

PROBLEM TO BE SOLVED: To solve problems in a summary formed from a part of a sentence which is not always objectively expressing the document content and information on a site of a document, is sometimes made with a long summary, and is possibly made with the same summary for plural documents in the past. SOLUTION: An anchor character string extracting means 11 extracts a URL of a link mate document and an anchor character string from an assembly of object documents stored in a document assembly storage part 21. A document type discriminating means 12 discriminates a document type of a link source document. A link relation discriminating means 13 discriminates the link relation between the link source document and a summarizing object document. A summary character string deciding means 14 imparts a score to respective anchor character strings by referring to a score information storage part 22 for prestoring the score for indicating propriety as a summary of the anchor character strings on the basis of an appearing frequency of the anchor character strings, the document type of the link source document, and the link relationship between the link source document and the summarizing object document, and summarizes the anchor character strings having the highest total score.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2002－163276
（P2002－163276A）
(43)公開日　平成14年６月７日(2002.6.7)

(51)Int.Cl.[7]		識別記号	FI		テーマコード（参考）
Ｇ０６Ｆ	17/30	２２０	Ｇ０６Ｆ　17/30	２２０Ａ	５Ｂ０７５
		１７０		１７０Ａ	５Ｂ０８２
		３４０		３４０Ｂ	
		４１９		４１９Ｂ	
	12/00	５４６	12/00	５４６Ｂ	

審査請求　未請求　請求項の数９　ＯＬ　（全 11 頁）　最終頁に続く

(21)出願番号　特願2000－358808(P2000－358808)

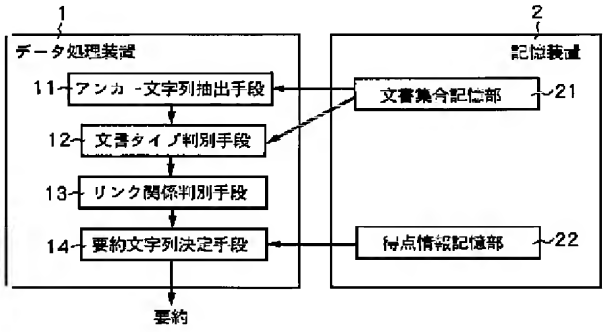
(22)出願日　平成12年11月27日(2000.11.27)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　赤峯　享
東京都港区芝５丁目７番１号　日本電気株式会社内
(72)発明者　杉浦　淳
東京都港区芝５丁目７番１号　日本電気株式会社内
(74)代理人　100086235
弁理士　松浦　兼行

Ｆターム(参考)　5B075 ND03 ND36 NS01 PQ02 PR04
5B082 AA01 BA09 EA04 GC04

(54)【発明の名称】　文書要約システム及び文書要約方法

(57)【要約】
【課題】　従来は、文章の一部分から作成した要約は、必ずしも文書内容と文書が置かれているサイトの情報を客観的に表しておらず、また、長い要約を作成してしまうことがあり 更に複数の文書に同じ要約を与える可能性がある。
【解決手段】　アンカー文字列抽出手段１１は、文書集合記憶部２１に格納された対象文書の集合からリンク先文書のＵＲＬとアンカー文字列を抽出する。文書タイプ判別手段１２は、リンク元文書の文書タイプを判別する。リンク関係判別手段１３は、リンク元文書と要約対象文書とのリンク関係を判別する。要約文字列決定手段１４は、アンカー文字列の出現頻度、リンク元文書の文書タイプ、及びリンク元文書と要約対象文書とのリンク関係を基に、アンカー文字列の要約としての適切さを示す得点を予め記憶している得点情報記憶部２２を参照して、各アンカー文字列に得点を付与し、合計得点が最も高いアンカー文字列を要約とする。



【特許請求の範囲】
【請求項１】　ＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約システムであって、
要約対象となるＨＴＭＬ文書の集合を予め記憶している文書集合記憶部と、
アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点を予め記憶している得点情報記憶部と、
前記文書集合記憶部のＨＴＭＬ文書の集合からリンク元文書のアンカー文字列を抽出するアンカー文字列抽出手段と、
前記アンカー文字列抽出手段により抽出されたリンク元文書が、リンク集であるかどうかを前記文書集合記憶部のＨＴＭＬ文書の集合から判別する文書タイプ判別手段と、
前記アンカー文字列抽出手段により抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、前記文書タイプ判別手段により判別された判別結果に基づき、前記得点情報記憶部に記憶されている得点情報を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定する要約文字列決定手段とを有することを特徴とする文書要約システム。
【請求項２】　ＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約システムであって、
要約対象となるＨＴＭＬ文書の集合を予め記憶している文書集合記憶部と、
アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点と、リンク元文書と要約対象文書とのリンク関係による要約としての適切さの得点とを予め記憶している得点情報記憶部と、
前記文書集合記憶部のＨＴＭＬ文書の集合からリンク元文書のアンカー文字列を抽出するアンカー文字列抽出手段と、
前記アンカー文字列抽出手段により抽出されたリンク元文書が、リンク集であるかどうかを前記文書集合記憶部のＨＴＭＬ文書の集合から判別する文書タイプ判別手段と、
前記アンカー文字列抽出手段により抽出されたリンク元文書と要約対象文書の関係を判別するリンク関係判別手段と、
前記アンカー文字列抽出手段により抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、前記文書タイプ判別手段により判別された判別結果と、前記リンク関係判別手段により判別されたリンク関係とに基づき、前記得点情報記憶部に記憶されている得点情報を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定する要約文字列決定手段とを有することを特徴とする文書要約システム。
【請求項３】　前記文書集合記憶部のＨＴＭＬ文書の集合を解析して、要約対象文書が属するサイトの代表文書とその代表文書の要約を取得する代表文書取得手段と、前記要約文字列決定手段により決定された要約対象文書の要約と同じ要約の文書が複数存在した場合、前記代表文書取得手段で取得した代表文書の要約と前記要約対象文書の要約とを連結して新たな要約として出力し、前記要約文字列決定手段により決定された要約対象文書の要約と同じ要約の文書が複数存在しない場合は、前記要約文字列決定手段により決定された要約対象文書の要約を出力する要約合成手段とを更に有することを特徴とする請求項１又は２記載の文書要約システム。
【請求項４】　前記要約文字列決定手段は、前記アンカー文字列抽出手段により抽出されたリンク元文書のアンカー文字列を単語に分割し、分割した単語の出現サイト数を数え、出現サイト数が多い方から順に前記出現頻度の順位を付け、前記得点情報記憶部に記憶されている得点情報を参照して前記順位の高いものほど出現頻度が多いとして高い得点を付与することを特徴とする請求項１乃至３のうちいずれか一項記載の文書要約システム。
【請求項５】　前記リンク関係判別手段は、前記要約対象文書のＵＲＬと前記アンカー文字列抽出手段により抽出されたリンク元文書のＵＲＬとを比較して、該リンク元文書が外部サイト文書、同一サイトの上位ディレクトリである上位文書、同一サイトの下位ディレクトリである下位文書、同一ＵＲＬの自文書、及びその他不明文書のいずれかとして前記リンク関係を判別し、前記得点情報記憶部は、前記外部サイト文書に対して最も高く、前記下位文書に対して最も低い得点情報を記憶していることを特徴とする請求項２記載の文書要約システム。
【請求項６】　ＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約方法であって、
ＨＴＭＬ文書の集合からリンク元文書のアンカー文字列を抽出する第１のステップと、
前記第１のステップにより抽出されたリンク元文書が、リンク集であるかどうかを前記ＨＴＭＬ文書の集合から判別する第２のステップと、
前記第１のステップで抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、前記第２のステップで判別された文書タイプ判別結果に基づき、アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点を予め記憶している得点情報記憶部を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定する第３のステップとを含むことを特徴とする文書要約方法。
【請求項７】　ＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約方法

であって、
ＨＴＭＬ文書の集合からリンク元文書のアンカー文字列を抽出する第１のステップと、
前記第１のステップにより抽出されたリンク元文書が、リンク集であるかどうかを前記ＨＴＭＬ文書の集合から判別する第２のステップと、
前記第１のステップにより抽出されたリンク元文書と要約対象文書のリンク関係を判別する第３のステップと、
前記第１のステップで抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、前記第２のステップで判別された文書タイプ判別結果と、前記第３のステップで判別されたリンク関係とに基づき、アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点と、リンク元文書と要約対象文書とのリンク関係による要約としての適切さの得点とを予め記憶している得点情報記憶部を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定する第４のステップとを含むことを特徴とする文書要約方法。
【請求項８】　前記ＨＴＭＬ文書の集合を解析して、要約対象文書が属するサイトの代表文書とその代表文書の要約を取得する第５のステップと、前記第４のステップにより決定された要約対象文書の要約と同じ要約の文書が複数存在した場合、前記第５のステップで取得した代表文書の要約と前記要約対象文書の要約とを連結して新たな要約として出力し、前記第４のステップにより決定された要約対象文書の要約と同じ要約の文書が複数存在しない場合は、前記第４のステップにより決定された要約対象文書の要約を出力する第６のステップとを更に有することを特徴とする請求項６又は７記載の文書要約方法。
【請求項９】　前記第４のステップは、前記第１のステップで抽出されたリンク元文書のアンカー文字列を単語に分割し、分割した単語の出現サイト数を数え、出現サイト数が多い方から順に前記出現頻度の順位を付け、前記得点情報記憶部に記憶されている得点情報を参照して前記順位の高いものほど出現頻度が多いとして高い得点を付与することを特徴とする請求項６乃至８のうちいずれか一項記載の文書要約方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は文書要約システム及び文書要約方法に係り、特にハイパーテキストマークアップランゲッジ（ＨＴＭＬ：Hyper Text Markup Language）文書の集合を検索する際に、検索結果として表示するための文書要約を作成する文書要約システム及び文書要約方法に関する。
【０００２】
【従来の技術】近年、インターネットの普及により、ＨＴＭＬ文書の数は膨大になり、膨大なＨＴＭＬ文書の中から利用者が必要とする文書を見つけるための手段として、検索エンジンが利用されている。
【０００３】検索エンジンは、利用者が入力したキーワードとマッチした複数の文書の要約を検索結果として表示する。検索エンジンの利用者は、その要約を基に実際にその文書にアクセスする価値があるかどうかの判別を行って、価値のある文書のみをアクセスする。従って、検索エンジンの利用者が、効率的に文書を見つけるためには、文書内容と文書が置かれているサイトの情報を客観的に表した要約の出来が重要になる。
【０００４】従来、この検索結果の文書要約としては、文書のタイトル、文書中の重要語や文書構造を基に文書の一部分を抽出した要約を使用している。例えば、ＨＴＭＬタグ情報と単語の出現頻度を利用して、要約文として適切なものを自動抽出する文書検索装置が従来知られている（特開平１０－３０７８３７号公報）。この従来の文書検索装置では、インターネット上に存在するワールドワイドウェブ（ＷＷＷ：World Wide Web）データの多数のユニフォームリソースロケイター（ＵＲＬ：Uniform Resource Locator）を保持するＵＲＬ記憶手段と、検索要求を入力するための入力手段と、ＵＲＬ記憶手段内に保持されているＵＲＬの検索を行う検索手段をもつ検索装置において、ＵＲＬによって指定されるＨＴＭＬデータに対して、ＨＴＭＬデータをインターネット上から取得し、そのＨＴＭＬデータ内の句読点とＨＴＭＬのタグの認識を行い、ＨＴＭＬデータ内に含まれている文章を抽出し、その文章の中から、要約文として適当なものを自動的に選択し、文章でＷＷＷデータの内容を知るようにしたものである。
【０００５】また、他の従来の文書検索装置として、ハイパーリンクの構造を利用した検索結果として、リンク元の文書のアンカー文字列を参照するようにした検索装置も文献１（1997年７月、情報処理学会研究報告VOL.99,NO.57(FI-55　DD-19)、p.73-80、「ハイパーリンクの構造を利用した検索結果の選択手法」）により開示されている。
【０００６】
【発明が解決しようとする課題】しかるに、上記の従来の文書検索装置は、それぞれ以下の課題を有している。第１の課題は、文章の一部分から作成した要約は、必ずしも文書内容と文書が置かれているサイトの情報を客観的に表していないということである。その原因は、文書内に検索結果の要約として適切な個所があるとは限らないためである。
【０００７】例えば、論文等では文書内容を的確に表すタイトルに関してさえも、ＨＴＭＬ文書では、タイトルを記述していない文書や、「新規に作成した文書」のように文書の要約としては意味のないタイトルを記述した文書が存在する。更に、検索エンジンでヒットし易くすることを目的に、文書内容とは無関係な人気キーワード

を文書中に故意にちりばめた文書も存在する。

【０００８】第２の課題は、検索結果として一度に表示できる文書数が１文書だけというような長い要約を作成してしまうことがあるということである。その原因は、複数の要約の候補から適切な要約を選択する手段が与えられていないためである。

【０００９】例えば、後者の文献１記載の従来の文書検索装置では、複数あるアンカー文字列から適切なものを選択する方法が記述されていない。すべてのアンカー文字列を表示すると、要約として不適切なアンカー文字列を含む長い文章を表示することになり、検索結果として一度に表示できる文書数が限られてしまう。このことは、携帯電話等の画面の大きさが限られた端末を利用して、文書検索を行う際に特に問題となる。

【００１０】第３の課題は、複数の文書に同じ要約を与える可能性があることである。その原因は、要約作成時に他の文書の要約と比較を行っていないためである。

【００１１】例えば、「サッカー」のことを記述した２つの文書があった場合、どちらの文書の要約もそれぞれ単独の要約として「サッカー」が適切であるとしても、検索結果としてどちらの文書も「サッカー」として表示されてしまうと、利用者はどちらの文書がより自分が必要とするかの判断ができない。

【００１２】本発明は以上の点に鑑みなされたもので、文書内の文字列だけでなく、リンク元文書のアンカー文字列も要約候補の文字列とすることで、客観的な要約を作成し得る文書要約システム及び文書要約方法を提供することを目的とする。

【００１３】また、本発明の他の目的は、複数の観点からアンカー文字列の要約としての適切さを判断し、最も適切なアンカー文字列を選択することで、必要最小限の短い要約を作成し得る文書要約システム及び文書要約方法を提供することにある。

【００１４】更に、本発明の他の目的は、検索結果として表示した際に、他の文書の要約と区別できる要約を作成し得る文書要約システム及び文書要約方法を提供することにある。

【００１５】

【課題を解決するための手段】上記の第１の目的を達成するため、第１の発明のＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約システムは、要約対象となるＨＴＭＬ文書の集合を予め記憶している文書集合記憶部と、アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点を予め記憶している得点情報記憶部と、文書集合記憶部のＨＴＭＬ文書の集合からリンク元文書のアンカー文字列を抽出するアンカー文字列抽出手段と、アンカー文字列抽出手段により抽出されたリンク元文書が、リンク集であるかどうかを文書集合記憶部のＨＴＭＬ文書の集合から判別する文書タイプ判別手段と、アンカー文字列抽出手段により抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、文書タイプ判別手段により判別された判別結果に基づき、得点情報記憶部に記憶されている得点情報を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定する要約文字列決定手段とを有する構成としたものである。

【００１６】また、上記の第１の目的を達成するため、第２の発明のＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約方法は、ＨＴＭＬ文書の集合からリンク元文書のアンカー文字列を抽出する第１のステップと、第１のステップにより抽出されたリンク元文書が、リンク集であるかどうかをＨＴＭＬ文書の集合から判別する第２のステップと、第１のステップで抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、第２のステップで判別された文書タイプ判別結果に基づき、アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点を予め記憶している得点情報記憶部を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定する第３のステップとを含むことを特徴とする。

【００１７】上記の第１及び第２の発明では、ＨＴＭＬ文書の集合から抽出したリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と文書タイプ判別結果に基づき、得点情報記憶部を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定するようにしたため、文書内の文字列だけでなく、リンク元文書のアンカー文字列も要約候補の文字列とすることができ、第１の目的を達成することができる。

【００１８】また、上記の第２の目的を達成するため、第３の発明のＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約システムは、上記の第１の発明における得点情報記憶部に、要約対象となるＨＴＭＬ文書の集合を予め記憶している文書集合記憶部と、アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点と、リンク元文書と要約対象文書とのリンク関係による要約としての適切さの得点とを予め記憶すると共に、アンカー文字列抽出手段により抽出されたリンク元文書と要約対象文書の関係を判別するリンク関係判別手段を設け、更に、上記の第１の発明における要約文字列決定手段を、アンカー文字列抽出手段により抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、文書タイプ判別手段により判別された判別結果と、リンク関係判別手段により判別されたリンク関係とに基づき、得点情報記憶部に記憶されている得点情報を参照して得点を付与

し、合計得点の最も高いアンカー文字列を要約として決定する構成としたものである。

【００１９】また、上記の第２の目的を達成するため、第４の発明のＨＴＭＬ文書の集合を検索する際に、検索結果として表示する文書要約を作成する文書要約方法は、ＨＴＭＬ文書の集合からリンク元文書のアンカー文字列を抽出する第１のステップと、第１のステップにより抽出されたリンク元文書が、リンク集であるかどうかをＨＴＭＬ文書の集合から判別する第２のステップと、第１のステップにより抽出されたリンク元文書と要約対象文書のリンク関係を判別する第３のステップと、第１のステップで抽出されたリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と、第２のステップで判別された文書タイプ判別結果と、第３のステップで判別されたリンク関係とに基づき、アンカー文字列の出現頻度による要約としての適切さの得点と、リンク元文書の文書タイプによる要約としての適切さの得点と、リンク元文書と要約対象文書とのリンク関係による要約としての適切さの得点とを予め記憶している得点情報記憶部を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定する第４のステップとを含むことを特徴とする。

【００２０】上記の第３及び第４の発明では、ＨＴＭＬ文書の集合から抽出したリンク元文書のアンカー文字列毎に、そのアンカー文字列の出現頻度と文書タイプ判別結果とリンク関係とに基づき、得点情報記憶部を参照して得点を付与し、合計得点の最も高いアンカー文字列を要約として決定するようにしたため、複数の観点からアンカー文字列の要約としての適切さを判断し、最も適切なアンカー文字列を選択することができ、必要最小限の短い要約を作成するという第２の目的を達成することができる。

【００２１】更に、上記の第３の目的を達成するため、第５の発明の文書要約システムは、第３の発明に加えて、文書集合記憶部のＨＴＭＬ文書の集合を解析して、要約対象文書が属するサイトの代表文書とその代表文書の要約を取得する代表文書取得手段と、要約文字列決定手段により決定された要約対象文書の要約と同じ要約の文書が複数存在した場合、代表文書取得手段で取得した代表文書の要約と要約対象文書の要約とを連結して新たな要約として出力し、要約文字列決定手段により決定された要約対象文書の要約と同じ要約の文書が複数存在しない場合は、要約文字列決定手段により決定された要約対象文書の要約を出力する要約合成手段とを更に有する構成としたものである。

【００２２】更に、上記の第３の目的を達成するため、第６の発明の文書要約方法は、第４の発明に加えて、ＨＴＭＬ文書の集合を解析して、要約対象文書が属するサイトの代表文書とその代表文書の要約を取得する第５のステップと、第４のステップにより決定された要約対象文書の要約と同じ要約の文書が複数存在した場合、第５のステップで取得した代表文書の要約と要約対象文書の要約とを連結して新たな要約として出力し、第４のステップにより決定された要約対象文書の要約と同じ要約の文書が複数存在しない場合は、第４のステップにより決定された要約対象文書の要約を出力する第６のステップとを更に有することを特徴とする。

【００２３】上記の第５及び第６の発明では、要約対象文書の要約と同じ要約の文書が複数存在した場合、要約対象文書が属するサイトの代表文書の要約と要約対象文書の要約とを連結して新たな要約として出力するようにしたため、検索結果として表示した際に、他の文書の要約と区別できる要約を作成できるという第３の目的を達成することができる。

【００２４】ここで、第１、第３及び第５の発明において、要約文字列決定手段は、アンカー文字列抽出手段により抽出されたリンク元文書のアンカー文字列を単語に分割し、分割した単語の出現サイト数を数え、出現サイト数が多い方から順に出現頻度の順位を付け、得点情報記憶部に記憶されている得点情報を参照して順位の高いものほど出現頻度が多いとして高い得点を付与することを特徴とする。

【００２５】また、第２、第４及び第６の発明において、第４のステップは、第１のステップで抽出されたリンク元文書のアンカー文字列を単語に分割し、分割した単語の出現サイト数を数え、出現サイト数が多い方から順に出現頻度の順位を付け、得点情報記憶部に記憶されている得点情報を参照して順位の高いものほど出現頻度が多いとして高い得点を付与することを特徴とする。これにより、要約としてより適切な得点を出現頻度から得ることができる。

【００２６】

【発明の実施の形態】（第１の実施の形態）次に、本発明の第１の実施の形態について図面と共に説明する。図１は本発明になる文書要約システムの第１の実施の形態のブロック図を示す。この実施の形態は、プログラム制御により動作するデータ処理装置１と、情報を記憶する記憶装置２とより構成される。

【００２７】記憶装置２は、文書集合記憶部２１と得点情報記憶部２２とを備えている。文書集合記憶部２１は、要約対象となるＨＴＭＬ文書の集合を予め記憶している。得点情報記憶部２２は、アンカー文字列の要約としての適切さを示す得点を予め記憶している。要約としての適切さを示す得点の例としては、アンカー文字列の出現頻度（出現サイト数）による得点、リンク元文書（被リンク先文書）の文書タイプがリンク集であるか否かによる得点、リンク元文書（被リンク先文書）と要約対象文書とのリンク関係による得点などがある。

【００２８】データ処理装置１は、アンカー文字列抽出手段１１、文書タイプ判別手段１２、リンク関係判別手

段１３及び要約文字列決定手段１３を備えている。アンカー文字列抽出手段１１は、文書集合記憶部２１に格納された対象文書の集合からリンク先文書のＵＲＬとアンカー文字列を抽出する。更に、アンカー文字列抽出手段１１は、抽出した結果をリンク元文書ＵＲＬとアンカー文字列の対応を示す表に変換し、要約対象文書毎にまとめる。

【００２９】文書タイプ判別手段１２は、リンク元文書の文書タイプを判別し、判別した文書タイプをアンカー文字列抽出手段１１が作成した表に追加する。文書タイプの例としては、リンク集がある。リンク関係判別手段１３は、リンク元文書と要約対象文書とのリンク関係を判別し、判別したそのリンク関係をアンカー文字列抽出手段１１が作成した表に追加する。リンク関係の例としては、外部サイト文書、上位文書、下位文書、自文書、及びその他・不明文書とがある。

【００３０】要約文字列決定手段１４は、アンカー文字列の出現頻度、リンク元文書の文書タイプ、及びリンク元文書と要約対象文書とのリンク関係を基に、得点情報記憶部２２の得点情報を参照して、各アンカー文字列に得点を付与し、合計得点が最も高いアンカー文字列を要約とする。

【００３１】次に、図２のフローチャートを併せ参照して図１の実施の形態の動作について詳細に説明する。まず、アンカー文字列抽出手段１１は、文書集合記憶部２１に格納された対象文書の集合を入力として受け、その入力文書からリンク先文書ＵＲＬと対応するアンカー文字列を抽出し、抽出した結果をリンク元文書ＵＲＬとアンカー文字列の対応を示す表に変換し、要約対象文字毎にまとめる（図２のステップＳ１１）。

【００３２】次に、文書タイプ判別手段１２は、被リンク先文書の文書タイプがリンク集であるかを判別し、アンカー文字列抽出手段１１が作成した表に文書タイプを追加する（図２のステップＳ１２）。次に、リンク関係判別手段１３は、リンク先文書と要約対象文書のリンク関係を判別する（図２のステップＳ１３）。

【００３３】次に、要約文字列決定手段１４は、アンカー文字列の出現頻度、リンク元文書の文書タイプ、及びリンク関係の情報を基に、得点情報記憶部２２の得点情報を参照し、各アンカー文字列に参照して得た得点を付与し（図２のステップＳ１４）、合計得点が最も高いアンカー文字列を要約として出力する（図２のステップＳ１５）。

【００３４】次に、本実施の形態の効果について説明する。本実施の形態では、要約を作成するのに、リンク元文書のアンカー文字列を利用している。そのため、文書内容と文書が置かれているサイトの情報を客観的に表した要約の作成が可能である。また、本実施の形態では、アンカー文字列の出現頻度、リンク元文書の文書タイプ、及びリンク元文書と対象文書のリンク関係という複数の観点から、複数のアンカー文字列の中で最も高い得点のアンカー文字列のみを選択しているため、適切な短い要約を作成することができる。

【００３５】（第２の実施の形態）図３は本発明になる文書要約システムの第２の実施の形態のブロック図を示す。同図中、図１と同一構成部分には同一符号を付してある。この第２の実施の形態は、プログラム制御により動作するデータ処理装置３が、図１に示したデータ処理装置１の構成に加え、代表文書取得手段３１と要約合成手段３２とを備える点で異なる。

【００３６】代表文書取得手段３１は、文書集合記憶部２１の文書集合を解析して、対象文書のサイトの代表頁を取得する。代表文書は、文献２（2000年1月、情報処理学会研究報告VOL.2000.NO.10（DS-20-2）p.9-16、サイテーション・エンジン、「リンク解析を用いたWWW検索ランキングシステム」）に記載されている代表頁と同じものであり、この文献２に開示された方法で代表文書を取得可能である。

【００３７】要約合成手段３２は、複数の文書に同じ要約が存在した場合、代表文書取得手段３１で取得した代表文書の要約と対象文書の要約を連結したものを要約として出力する。

【００３８】次に、図４のフローチャートを併せ参照して図３の実施の形態の動作について詳細に説明する。図４中、図２と同一処理ステップには同一符号を付し、その説明を省略する。図３の要約合成手段３２は、要約文字列決定手段１４により決定された対象要約の中に、同一の要約の文書が存在するかどうか調べ（図４のステップＳ２１）、同一の要約の文書が存在した場合、代表文書取得手段３１で取得した代表文書の要約を受け（図４のステップＳ２２）、この代表文書の要約と上記の対象要約とを連結したものを要約として（図４のステップＳ２３）、出力する（図４のステップＳ２４）。

【００３９】一方、要約合成手段３２は、ステップＳ２１で同一の要約の文書が存在しないと判断した場合は、要約文字列決定手段１４により決定された対象要約をそのまま要約として出力する出力する（図４のステップＳ２４）。

【００４０】次に、本実施の形態の効果について説明する。本実施の形態では、一旦要約候補を作成した後、同じ要約の文書が存在するかどうかチェックし、同じ要約の文書が存在するときには、代表文書の要約と対象要約とを連結したものを要約として出力するようにしたため、複数の文書が同じものになることを防止することができ、また、他の文書と区別可能な要約を作成することができる。

【００４１】

【実施例】次に、本発明の第１の実施例を図面と共に説明する。本実施例は第１の実施の形態に対応した実施例である。本実施例は、データ処理装置１としてパーソナ

ルコンピュータ、記憶装置２として磁気ディスク記憶装置とを備えている。パーソナルコンピュータは、アンカー文字列抽出手段１１、文書タイプ判別手段１２、リンク関係判別手段１３、要約文字列決定手段１４を有しており、磁気ディスク記憶装置には、文書集合記憶部２１と得点情報記憶部２２を有している。

【００４２】図５は対象文書集合中の文書の一例を示す。アンカー文字列抽出手段１１は、図５のＵＲＬがhttp://aa.bb/xxの文書から図７（Ａ）に示すようなリンク先ＵＲＬ「http://aa.bb/xx/b」とアンカー文字列「野球」の対応と、リンク先ＵＲＬ「http://aa.bb/xx/s」とアンカー文字列「サッカー」の対応とを抽出する。

【００４３】図７（Ａ）の場合、タグで明示的に囲まれた文字列のみをアンカー文字列として抽出しているが、例えば図５の文章からタグの前後の文字列も合わせてアンカー文字列として抽出することや、タイトルを自文書へのアンカー文字列として抽出することで、図７（Ｂ）に示す文字列もアンカー文字列も抽出することができる。また、本実施例ではアンカー文字列として名詞句のみを扱っているが、文をアンカー文字列として抽出することもできる。

【００４４】次に、アンカー文字列抽出手段１１は、抽出した対応をリンク元文書ＵＲＬとアンカー文字列の対応に変換し、各要約対象文書に対して対応表を作成する。図８に文書「http://aa.bb/xx/s」に対して、アンカー文字列抽出手段１１が作成したリンク元文書ＵＲＬとアンカー文字列の対応表の例を示す。この対応表のリンク元文書ＵＲＬ「http://aa.bb/xx」とアンカー文字列「サッカー」の対応は、図７（Ａ）のリンク先文書ＵＲＬ「http://aa.bb/xx/s」とアンカー文字列「サッカー」の対応を変換したものである。

【００４５】文書タイプ判別手段１２は、例えば文書が３つ以上の異なる外部サイトへのリンクを持っている場合、その文書をリンク集と判定する。図９はリンク集である文書の一例を示す。図９の文書「http://xx.hh/aa」は、自サイトが「xx.hh」であり、外部サイト「aa.bb」、「xx.yy」及び「xx.zz」へのリンクを持っている。従って、３つ以上の異なる外部サイトへのリンクを持っているので、「http://xx.hh/aa」は、リンク集であると判定する。

【００４６】なお、本実施例では、文書タイプの判別方法として、外部サイトへのリンク数による判別方法を述べたが、他にも文献３（1999年、情報処理学会研究報告VOL.99,NO.20(FI-53) p.9-16、「文書タイプ分類による問題解決向きWWW検索システムの開発と評価」）に示されたような、文書内に「リンク集」という単語が存在することと外部サイトへのリンクが存在することとを組み合わせて、文書タイプを総合的に判定する方法もあり、ここで述べた方法に限定されない。

【００４７】リンク関係判別手段１３は、文書ＵＲＬと被リンク先の文書ＵＲＬを比較して、リンク元の文書が外部サイト文書か、上位文書か、下位文書か、自文書か、その他・不明文書かを判別する。図１０は図８の対応表に文書タイプ判別判別手段１２が付与した文書タイプの項目と、リンク関係判別手段１３が付与したリンク関係の項目を追加した対応表の一例を示す。

【００４８】図１０に示すように、文書「http://aa.bb/xx/yy」を基準にした場合、文書「http://xx.hh/aa」や文書「http://gg.hh/bb」はそれぞれサイトが異なるので、外部サイト文書であり、文書「http://aa.bb/xx」は同一サイトで上位のディレクトリなので、上位文書であり、文書「http://aa.bb/xx/yy/w1」及び文書「http://aa.bb/xx/yy/w2」は、それぞれ同一サイトで下位のディレクトリなので下位文書であり、文書「http://aa.bb/xx/yy」は同じＵＲＬなので自文書である。

【００４９】要約文字列決定手段１４は、アンカー文字列を単語に分割し、分割した単語の出現サイト数を数え、より多くのサイトに出現するアンカー文字列が上位になるように順位をつける。図１０の文書「http://aa.hh/xx/yy」では、アンカー文字列として、「最新情報」、「戻る」、「サッカー」、「Ｊリーグ情報」、「サッカー速報」が存在する。

【００５０】例えば、「最新情報」は、「最新」と「情報」の２単語に分解され、それぞれの単語が出現するサイトは、aa.bbだけの１サイトであり、「サッカー速報」は、「サッカー」と「速報」の２単語に分解され、それぞれの単語が出現するサイトは、aa.bb、xx.hh、gg.hhの３サイトである。図１１は、図１０に示した各アンカー文字列に対し、分割した単語と、分割した単語が出現するサイトと、出現サイト数と、出現サイト数による順位の例を示す。図１１に示すように、「サッカー速報」が出現サイト数３で１位に、「Ｊリーグ速報」と「サッカー」が出現サイト数２で２位に、「最新情報」と「戻る」が出現サイト数１で４位になる。

【００５１】更に、要約文字列決定手段１４は得点情報記憶部２２に予め記憶している得点情報を参照して、出現サイト数による順位、リンク元文書の文書タイプ、リンク元文書と要約対象文書のリンク関係による得点を与え、最も合計得点の高いアンカー文字列を要約とする。図６は得点情報記憶部２２の得点情報の一例を示す。ここでは、アンカー文字列の出現頻度の最も高いものを１０点とし、以下、文字列の出現頻度の順に５点、３点、１点としている。また、文書タイプがリンク集であれば１０点とする。更に、リンク関係では外部サイト文書が１０点、上位文書及び自文書がそれぞれ５点、下位文書が０点、その他・不明文書が３点としている。

【００５２】なお、同じアンカー文字列に対してリンク元文書の文書タイプやリンク元文書と要約対象文書とのリンク関係が複数ある場合は、高い方の得点をそのアンカー文字列の得点とする。

【００５３】図１０と図１１の表の値に対して、図6の得点情報を参照した場合の得点を図１２に示す。図１２に示すように、「最新情報」は、出現サイト数の順位が４位なので、出現サイト数による得点は１点、文書タイプがリンク集でないので文書タイプによる得点は０点、リンク関係は自文書なのでリンク関係による得点は５点となり、合計得点は６点となる。

【００５４】また、「Ｊリーグ速報」は出現サイト数の順位が２位なので、出現サイト数による得点は５点、文書タイプがリンク集なので文書タイプによる得点は１０点、リンク関係は外部サイト文書なのでリンク関係による得点は１０点となり、合計得点２５点となる。図１２の例では、最も合計得点の高いアンカー文字列の「Ｊリーグ速報」を要約として選択する。

【００５５】次に、本発明の第２の実施例を、図面を参照して説明する。本実施例は、図３に示した第２の実施の形態に対応するものである。本実施例は、第１の実施例と構成を同じとするが、パーソナルコンピュータの中央演算装置が代表文書取得手段３１及び要約合成手段３２としても機能する点で第１の実施例と異なる。

【００５６】今、第１の実施例と同じ方法で要約文字列決定手段１４で文書「http://aa.bb/xx/yy」に対して「Ｊリーグ速報」が要約として選択されたとする。また、文書「http://bb.aa/xx/yy」に対しても、「Ｊリーグ速報」が要約として選択されているとする。

【００５７】要約合成手段３２は、文書「http://aa.bb/xx/yy」の要約「Ｊリーグ速報」と同じ要約が存在するかを調べる。本実施例では、同じ要約が文書「http://bb.aa/xx/yy」に存在するため、代表文書取得手段３１が文書「http://aa.bb/xx/yy」の代表文書とその代表文書の要約を取得する。

【００５８】本実施例では、代表文書が「http://aa.bb/」でその要約が「Ａ新聞」であったとする。要約合成手段３２は、代表文書の要約の「Ａ新聞」と、対象文書の要約の「Ｊリーグ速報」を連結して「Ａ新聞Ｊリーグ速報」を要約として出力する。

【００５９】このように、本実施例では、同じ要約の文書が存在するときには、代表文書の要約と対象要約とを連結したものを要約として出力するようにしたため、複数の文書が同じものになることを防止することができ、また、他の文書と区別可能な要約を作成することができる。

【００６０】なお、本発明は以上の実施の形態及び実施例に限定されるものではなく、例えば、第１の実施の形態において、リンク関係判別手段１３は必ずしも有していなくてもよく、その場合は、要約文字列決定手段１４は、アンカー文字列の出現頻度、リンク元文書の文書タイプを基に、得点情報記憶部２２の得点情報を参照して、各アンカー文字列に得点を付与し、合計得点が最も高いアンカー文字列を要約とする。

【００６１】

【発明の効果】以上説明したように、本発明によれば、文書内の文字列だけでなく、リンク元文書のアンカー文字列も要約候補の文字列とすることにより、文書内容と文書が置かれているサイトの情報を客観的に表した要約の作成ができるため、検索エンジンの検索結果として、この要約が表示された場合、利用者は文書がアクセスする価値があるかどうかを容易に判別することができる。

【００６２】また、本発明によれば、複数の観点からアンカー文字列の要約としての適切さを判断し、最も適切なアンカー文字列を選択することにより、必要最小限の短い要約を作成することができるようにしたため、検索エンジンの検索結果としてこの要約を表示する場合、複数の検索結果を一画面に表示することができる。

【００６３】更に、本発明によれば、要約対象文書の要約と同じ要約の文書が複数存在した場合、要約対象文書が属するサイトの代表文書の要約と要約対象文書の要約とを連結して新たな要約として出力することで、検索結果として表示した際に、他の文書の要約と区別できる要約を作成できるようにしたため、検索エンジンの検索結果として、この要約が表示された場合、利用者は複数の文書を区別することができ、より適切な文書にアクセスすることができる。

【図面の簡単な説明】

【図１】本発明システムの第１の実施の形態のブロック図である。

【図２】図１の動作を説明する本発明方法の第１の実施の形態のフローチャートである。

【図３】本発明システムの第２の実施の形態のブロック図である。

【図４】図３の動作を説明する本発明方法の第２の実施の形態のフローチャートである。

【図５】本発明の第１の実施例のアンカー文字列である。

【図６】本発明の第１の実施例の得点情報記憶部の得点情報の一例である。

【図７】本発明の第１の実施例のアンカー文字列の各例である。

【図８】本発明の第１の実施例のアンカー文字列抽出部が作成する表の一例である。

【図９】本発明の第１の実施例のリンク集合の文書の一例である。

【図１０】本発明の第１の実施例のリンク元文書の文書タイプとリンク元文書と対象文書のリンク関係の一例を示す図である。

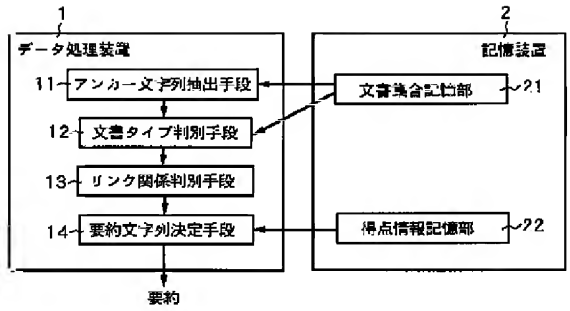
【図１１】本発明の第１の実施例のアンカー文字列の出現サイト数による順位付けを説明するための図である。

【図１２】本発明の第１の実施例の要約文字列決定手段の得点計算を説明するための図である。

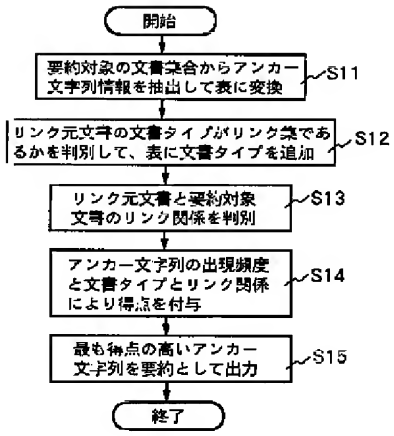
【符号の説明】

１、３　データ処理装置
２　記憶装置
１１　アンカー文字列抽出手段
１２　文書タイプ判別手段
１３　リンク関係判別手段
１４　要約文字列決定手段
２１　文書集合記憶部
２２　得点情報記憶部
３１　代表文書取得手段
３２　要約合成手段

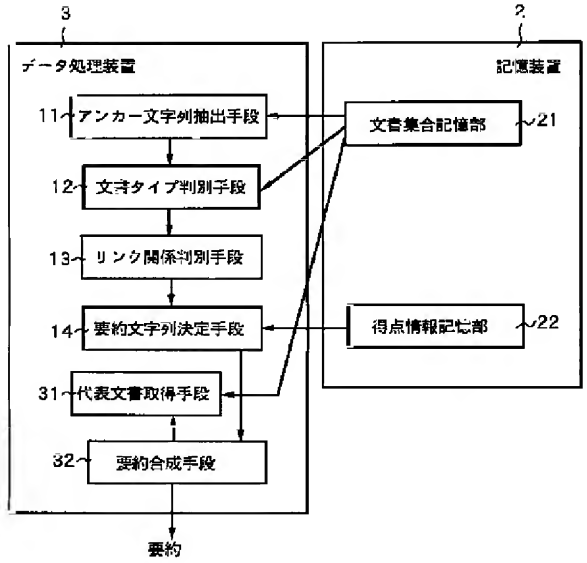
【図１】



【図２】



【図３】



【図５】

http ://aa.bb/xx

```
<TITLE>Aスポーツ</TITLE>
-------
<A HREF="http://aa.bb/xx/b>野球</A>
プロ野球情報
-------
<A HREF="http://aa.bb/xx/s>サッカー</A>
Jリーグ情報
```

【図６】

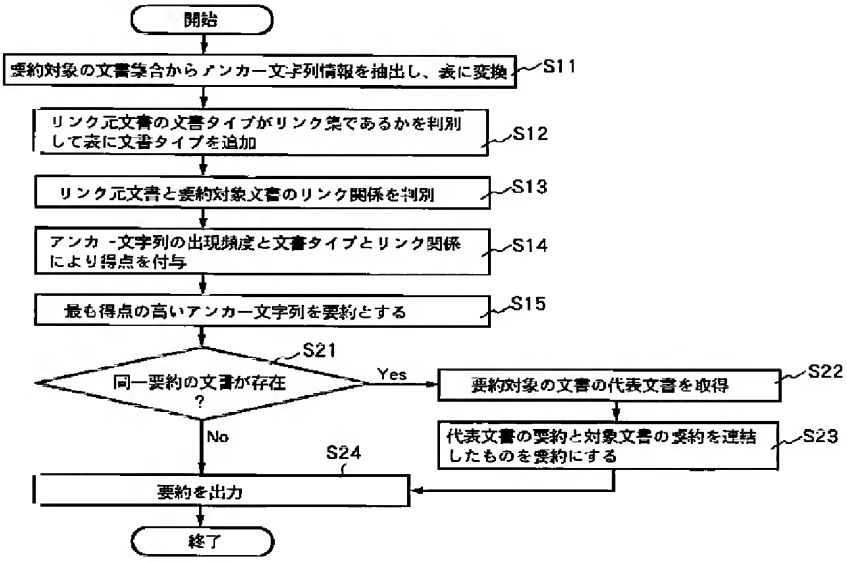
・アンカー文字列の出現頻度	
1位（最も多くのサイトに出現）	１０点
2位（次に多くのサイトに出現）	５点
3位（その次に多くのサイトに出現）	３点
3位以下（その他）	１点
・文書タイプ	
リンク集	１０点
・リンク関係	
外部サイト文書	１０点
上位文書	５点
自文書	５点
下位文書	０点
その他・不明文書	３点

【図８】

http ://aa.bb/xx/s

リンク元文書のＵＲＬ	アンカー文字列
http ://aa.bb/xx/s	最新情報
http ://aa.bb/xx/s/w1	戻る
http ://aa.bb/xx/s/w2	戻る
http ://aa.bb/xx	サッカー
http ://xx.hh/z1	Ｊリーグ速報
http ://gg.hh/z3	サッカー速報

【図4】



【図7】

(A)

http://aa.bb/xx

リンク先文書のURL	アンカー文字列
http://aa.bb/xx/b	野球
http://aa.bb/xx/s	サッカー

(B)

http://aa.bb/xx

リンク先文書のURL	アンカー文字列
http://aa.bb/xx	スポーツ
http://aa.bb/xx/b	野球
http://aa.bb/xx/b	プロ野球情報
http://aa.bb/xx/s	サッカー
http://aa.bb/xx/s	Jリーグ情報

【図9】

http://xx.hh/aa

<TITLE>スポーツ速報</TITLE>

リンク集

<A HREF="http://aa.bb/xx/ss">サッカー速報</A>

<A HREF="http://xx.yy/mm">高校野球速報</A>

<A HREF="http://xx.zz/nn">オリンピック速報</A>

【図10】

http: aa.bb xx yy

リンク元文書のURL	アンカー文字列	文書タイプ	リンク関係
http: aa.bb xx yy	最新情報		自文書
http: aa.bb xx yy w1	戻る		下位文書
http: aa.bb xx yy w2	戻る		下位文書
http: aa.bb xx	サッカー		上位文書
http: xx.hh aa	Jリーグ速報	リンク集	外部サイト文書
http: gg.hh bb	サッカー速報		外部サイト文書

【図11】

アンカー文字列	単語	出現サイト	出現サイト数	順位
最新情報	最新、情報	aa.bb	1	4
戻る	戻る	aa.bb	1	4
サッカー	サッカー	aa.bb、gg.hh	2	2
Jリーグ速報	Jリーグ、速報	xx.hh、gg.hh	2	2
サッカー速報	サッカー、速報	aa.bb、xx.hh、gg.hh	3	1

【図１２】

アンカー文字列	出現サイト数	文書タイプ	リンク関係	合計得点
最新情報	１点	０点	５点	６点
戻る	１点	０点	０点	１点
サッカー	５点	０点	５点	１０点
Ｊリーグ速報	５点	１０点	１０点	２５点
サッカー速報	１０点	０点	１０点	２０点

フロントページの続き

(51)Int.Cl.[7]	識別記号	ＦＩ		(参考)
Ｇ０６Ｆ　12/00	５４７	Ｇ０６Ｆ　12/00	５４７Ｈ	